Identification and Rejection of Outliers in Flood Frequency Estimation

Dr. Supon Boripun*

1. Introduction

In the design situation, engineers, planners, farmers, etc., need information of flood magnitude and frequency [1]. Frequently, outliers may appear in the flood series recorded for flood analysis [2]. Outliers are data points deviating from the general trend of the other data. Outliers may by caused be error in the data recording process, or, sometimes, unusual events of nature. Outliers often provide a difficult practical problem, especially, their exclusion may change the derived flood parameters considerably. This may lead to misleading interpretations of data analysis. The result of the analysis may be useless, or cause an unexpected disaster if the safety of the project is based on this result, such as in flood control or dam construction. Hence, it is very important to be careful about outliers. The procedure for identification and treatment of outliers in flood frequency estimation will be discussed in this paper.

2. Identification of Outliers

Outliers occur in flood data records occasionally, especially for first and second ranked flood events in a flood series [3]. These

*Lecturer, Department of Physics, Khon Kaen University, Khon Kaen 40002

extreme events can be treated if some historical information is available. The problem is whether these outliers can be included in the flood frequency analysis. To consider this matter, the identification of the outliers should be done first. The identification can be divided into three categories;

1. Cause of the occurrence. It is very important to establish the reason for this apparent anomaly, as the cause of the error is very important for a decision on its treatment.

2. Degree of prior belief. As the identification of the causes of the outliers is not always straightforward, this can be generally categorized into the degrees of confidence in the results.

3. Degree of evidence. Since it is difficult to decide whether the outlier should be excluded or not, evidence of its deviation from the sample data should be examined in order to make a decision.

These three categories will be discussed in detail as follows.

3. Cause of the Occurrence

The flood data should be istablished whether errors exist in the basic hydrologic data (especially the station rating curve) and the data sample should be checked to ensure that it is homogeneous [3]. A statistical evaluation of the outliers should be made and if the deviation of the outlier is identified as being statistically significant, the causes of error may be any of the following classes.

1. Recording error. This is an incorrect observation which should be carefully checked. The error may be due to an error in

the measuring equipment. To examine the local flood records, nearby flood records and the station rating curve will be of great help to justify the conclusion.

2. Resulting from extreme occurrence of the same phenomenon responsible for all other observations. This is a rare event that may occur during the recording of the flood data series. All of the data, including the outliers, resulted from the same cause, but for the outliers the events were extreme. For example, an extreme flood event of a specific catchment may result from the cyclone phenomena. There may be an abnormally large amount of rain which could also have been caused by the cyclone and these events may provide the outlier data.

3. Resulting from rare alternate phenomena. This is the most serious problem. Actually, the phenomena is different to that which caused other floods on record, and this provides very few observations. Thus any statistical conclusions based on these rare events are impossible to draw. The existence of this cause can only be ascertained from a thorough knowledge of the area and its climatological characteristics.

4. Degree of Prior Belief

The outliers may have resulted from any of those three classes described previously. An identification of these classes can be generally categorized as follows.

1. A strong prior belief that the outliers fall into one particular class.

2. A moderate prior belief that the outliers fall into one class, but there is still a suspicious that they may fall in to another class.

3. A weak prior belief that the outliers are believed to be in one of the classes but particular class cannot be justified.

5. Degree of Evidence

For verification of the outliers, the suspected observations should be omitted from the sample as a first approximation. The truncated data should be used for fitting of a frequency distribution. Usually, the 95% and 5% confidence limits should also be calculated and plotted. Then the outliers should be plotted in their respective positions according to the equation T = (N+1)/R; where T = return period; N = number of years recorded and R = rank of the flood in the series [3]. The evidence for the proposition that the suspected observation is an outlier can now be justified as the following categories [37].

1. Strong evidence. The observation will be categorized as strong evidence of being an outlier if it plots outside the confidence bands, and also plots outside the bands although it is considered as a flood at any reasonable recurrence interval (say within 1,000 years).

2. Moderate evidence. In this category the observation plots outside the confidence bands but not as far as the strong category. The plot will be in the bands at some reasonable recurrence intervals of the flood events. Theoretically, the chance is 1 in 20 for one side or 1 in 10 for both side of the confidence bands that an event lying outside is a member of the assumed frequency distribution.

3. Weak evidence. The observation plots within the confidence bands but there is still a distinct possibility that the observation resulted differently from the rest of the data. In this case the evidence is clearly weaker thaw the previous two categories.

6. Treatment of Outliers

The identified outliers will, then, be treated as the recommendation in the table below.

The term "omit" implies that the outlier should be discarded from the sample. The Institute of Engineers, Australia [3], suggested that on the basis of cause 3 although the suspected observation is omitted in the frequency analysis, the possibility of the extreme event resulting from the alternate phenomenon should be noted.

The term "modify probability" implies acceptance teat the suspected observation is a member of the population of all the other population. Its probability of exceedance must be inferred from the assumed frequency distribution, that is the recurrence interval is modified. There still, theoretically, remains a 1 in 10 chance of being wrong.

The term "modify magnitude" is for the case where some knowledge of the direction and magnitude of the suspected error in known.

The term "keep" shows that the suspected observation is accepted as a member of the sample population and there is no need to take any action.

All of the "keep" or "modify magnitude" outliers, together with the sample used in the first approximation, should be recomputed for the frequency distribution analysis. The decision to keep or modify should be re-evaluated if the new frequency distribution is very different from the first approximation.

Table 1. Guide for treatment of outliers [3],

CAUSE	EVIDENCE	PRIOR BELIEF		
		Strong	Moderate	weak
1 error in observation	Strong	Omit	Omit	Omit
	Moderate	Omit	Omit or modify magnitude	Modify probability or magnitude
	Weak	Omit	Keep or modify magnitude	Keep
2 rare event from the same population	Strong	Modify Probability	Omit or modify probability	Omit
	Moderate	Modify Probability	Modify Probability	Modify Probability
	Weak	Keep	Keep	Keep
3 Caused by rare alternate phenomenon	Stong	Omit	Omit	Omit
	Moderate	Omit	Omit or modify probability	Modify probability
	Weak	Omit	Omit or keep	keep

No further work should be done for the case of "omit" the the outlier, but to note the possibility of another extreme event in the case of cause 3 should be done.

The confidence bands will be recomputed for the case of "modify probability" but the first approximation of the frequency distribution must be left alone.

7. Conclusion

The outliers have a significant effect on the flood frequency analysis, in turn, this may cause difficulties in practical problems. Identification and treatment of the outliers are necessary and can be simply followed from the guide shown in Table 1. Some definitions and the procedure of identification concerning this matter should be recognized for the justification or rejection of the outliers.

8. Reference

1. Ward, R., [1978], Flood; A Geographical Perspective. Mcmillan, London.

2. Haan, C.T., [1977], Statistical Methods in Hydrology. The Iowa State Univ Press, Ames.

3. Anon [1977], Australian Rainfall and Runoff; Flood Analysis and Design, The Institute of Engineers, Australia.